OBSERVATORIO

EN LA ERA DE LA WEB DE LOS DATOS: PRIMERO DATOS ABIERTOS, DESPUÉS DATOS MASIVOS

In the age of the web of data: first open data, then big data

Tony Hernández-Pérez

Tony Hernández-Pérez es doctor en ciencias de la información y profesor del *Departamento de Biblioteconomía y Documentación* de la *Universidad Carlos III de Madrid* en donde dirige el programa de doctorado en documentación. Su labor docente e investigadora está ligada al grupo *TecnoDoc* incluyendo asignaturas de web social, gestión de contenidos web, metadatos, búsqueda y recuperación de información, e-learning y documentación periodística y audiovisual.
http://orcid.org/0000-0001-8404-9247

Universidad Carlos III de Madrid
Facultad de Humanidades, Comunicación y Documentación
C/ Madrid, 128. 28903 Getafe (Madrid), España
tony@bib.uc3m.es

Resumen

Repaso de los conceptos y las tecnologías asociadas a la evolución de una web de documentos a una web de datos. Papel que las bibliotecas públicas y académicas están jugando o pueden jugar respecto a los datos masivos, los datos abiertos y los datos abiertos vinculados. Se subraya la importancia estratégica que para el futuro de las bibliotecas pueden tener los datos abiertos y especialmente los datos abiertos vinculados.

Palabras clave

Datos masivos; Datos abiertos; Datos abiertos vinculados; Bibliotecas públicas; Bibliotecas académicas.

Abstract

Review of the concepts and technologies associated with the transition from a web of documents to a web of data. The role that public and academic libraries are playing or may play about big data, open data, and linked open data is described. The strategic importance of open data and linked open data (LOD) for the future of libraries is emphasized.

Keywords

Big data; Open data; Linked open data; LOD; Public libraries; Academic libraries.

Hernández-Pérez, Tony (2016). "En la era de la web de los datos: primero datos abiertos, después datos masivos". *El profesional de la información*, v. 25, n. 4, pp. 517-525.

http://dx.doi.org/10.3145/epi.2016.jul.01

1. Introducción

Las bibliotecas y los medios de comunicación viven tiempos de crisis. Se cierran medios de comunicación, se reducen o congelan presupuestos en las bibliotecas y se despide o no se renueva a muchos de sus profesionales. Ocurre lo obvio, como consecuencia de que las bibliotecas han perdido el monopolio de los libros como medio de transmisión de conocimiento y cultura. Y los medios de comunicación tradicionales el de la producción y distribución de noticias. En el mundo digital en que nos movemos ahora la información se consume principalmente a través de dispositivos móviles e internet (**Negredo**; **Vara-Miguel**; **Amoedo**, 2016) y ni la biblioteca ni los medios poseen ya en exclusiva el valor de la información como agente transformador de conocimiento y cultura.

Google y las redes sociales se han convertido en las principales puertas de acceso a la información. En este contexto de crisis y de sensación de pérdida de funciones, en el que

Artículo recibido el 14-07-2016

un *youtuber* puede tener más seguidores que un medio de comunicación, los periodistas y los bibliotecarios, con cada vez menos recursos, intentan "reinventarse" ampliando el espectro y el formato de noticias a publicar. En el caso de las bibliotecas, ampliando o creando nuevos servicios: apoyo a la publicación en revistas y a la petición de sexenios para investigadores, servicios de repositorios de investigación, apoyo a la alfabetización digital, al e-learning, etc. Son de alabar los esfuerzos por evolucionar, por innovar y casi por sobrevivir.

La Web, además, está pasando de ser un medio en el que publicar y compartir documentos —en el que hasta ahora nos sentíamos más o menos cómodos— a un medio en el que lo que se publica y comparte cada vez más son datos ya que incluso los documentos vienen marcados para poder tratarlos como datos. ¿No es paradójico? Después de pensar que pasábamos de la sociedad de la información a la sociedad del conocimiento regresamos a los orígenes de los procesos de automatización y parece que los datos vuelven a ser esenciales: *big data*, *data mining*, *open data*, *research data*, *linked data*, *open linked data*, *data science*, *data literacy*, *social data*, *data infrastructure*, *data privacy*..., son algunos de los términos frecuentes en la bibliografía científica de casi cualquier campo científico e incluso en los medios no especializados.

No, no se trata de ninguna paradoja sino de la evolución hacia la madurez de las tecnologías de la información: aumento y abaratamiento de la producción en formato digital, aumento de las capacidades personales y tecnológicas, de almacenamiento, de procesamiento, de acceso, de análisis, de enlace y de distribución..., que hace que ahora la pirámide del conocimiento de la que hablaba Russell Ackoff (**Rowley**, 2007) dato - información - conocimiento se cierre y sea realmente un ciclo. Y de la evolución de la Web como un medio para publicar documentos a un medio para compartir datos estructurados.

Naciones Unidas (2014) habla de una revolución de datos que tiene que ver con la gran cantidad de ellos que se generan y se almacenan cada día, y cuyo procesamiento puede servir como evidencia para la toma de decisiones bien fundamentadas. La *OCDE* (2015), el gobierno de EUA (**Obama**, 2009) y la *Comisión Europea* (2014) conceden la máxima importancia a la gestión de datos de todo tipo y la consideran una parte fundamental de la economía digital que regirá el mundo en los próximos años.

Figura 1. Nube de etiquetas relacionadas con los datos

La nueva fiebre del oro son los datos. De acuerdo con *IBM*:

> "cada día se crean 2,5 trillones de bytes de datos –supone que el 90% de los datos en el mundo de hoy se ha creado en los últimos dos años. Proceden de todas partes: sensores utilizados para recopilar información sobre el clima, mensajes a sitios de medios sociales, fotos y vídeos digitales que compartimos, registros de transacciones de compra o señales de GPS del teléfono móvil, por citar unos pocos" (*IBM*, 2016).

Todos estos datos conforman los "datos masivos" o *big data*.

Y en esa carrera por no perder el pie muchos se abrazan a los nuevos términos o conceptos que aparecen, como los citados antes. Uno de esos términos, de moda desde hace unos años, es *big data* (BD). Conste que no hay escepticismo respecto al mismo. No es una moda y está aquí para quedarse, sin duda.

> "No está tan claro que los profesionales de la información, bibliotecarios o comunicadores, tengan que tratar habitualmente con *big data*, más bien no"

Lo que ya no está tan claro es que los profesionales de la información, bibliotecarios o comunicadores, puedan considerar que los *big data* serán algo con lo que tratarán habitualmente, más bien no. Muy pocos podrán integrarse, ni siquiera a medio plazo, en equipos multidisciplinares que trabajen con datos masivos y habrá que distinguir entre *data scientist*, *data analyst*, otras ramas denominativas que seguro irán apareciendo.

2. Datos masivos o *big data* ¿Concierne hoy a las grandes bibliotecas?

Los datos masivos son fundamentalmente datos transaccionales, millones de búsquedas en tiempo real, millones de transacciones de compras o millones de posibles combinaciones químicas. Aunque no existe un estándar respecto a cuán grande debe ser un conjunto de datos, para considerarlo datos masivos y sólo a efectos didácticos se puede convenir que:

- Ficheros con datos de hasta 10 GB (decenas de miles de filas y columnas), procesables en *Excel* o *R* y manejables en la memoria de un solo ordenador pueden considerarse "datos pequeños".
- Ficheros de entre 10 GB y 1 TB (millones de páginas web) ya requieren para su manejo bases de datos algo especializadas y no funcionan bien en *Excel*, pero se pueden almacenar y trabajar desde un disco duro externo del ordenador.
- Ficheros mayores de 1 TB ya se consideran datos masivos, aunque cada vez se tiende a hablar más de petabytes o de exabytes. Para manejarlos normalmente se requieren bases de datos distribuidas y es necesario almacenarlos

SCIPEDIA
Register for free at https://www.scipedia.com to download the version without the watermark

en múltiples ordenadores que deben ser, por ejemplo, capaces de gestionar billones de clicks, como los que hacen decenas de miles de personas cuando juegan a *Candy Crush*.

A principios de los años 2000 los datos masivos se relacionaban con un modelo basado en las 3 V's:

- Volumen (en el sentido de que la cantidad de datos de los que se dispone exceden la capacidad de procesamiento de las bases de datos convencionales);
- Velocidad (en el sentido en que el tiempo en que se recopila y se procesa tal volumen de información es muy rápido, a veces incluso instantáneo); y
- Variedad (en el sentido de que las fuentes de las que proceden y los tipos de datos que se procesan pueden ser muy diferentes).

El *National Institute of Standards and Technology* (*NIST*, 2015) amplía su definición con una cuarta V, Variabilidad (los cambios que se producen en las características de los datos).

Al modelo se le han añadido algunas otras V's:

- Veracidad (en el sentido de que los datos deben ser fiables) o
- Valor (en el sentido de que la transformación de los datos en información y conocimiento es lo que ofrece una ventaja competitiva a las organizaciones que trabajan con estas masas ingentes de datos) pero sólo las primeras cuatro son reconocidas por *NIST*.

Trabajar con este tipo de datos requiere una arquitectura tecnológica especial para un eficiente almacenamiento, manipulación y análisis de los datos, una arquitectura muy vinculada a *Hadoop*[1], al aprendizaje automático y a una infraestructura con un sistema de ficheros diferente a las que conocemos, con algoritmos que funcionan de forma paralela y distribuida y con bases de datos *NoSQL*[2].

> "Sólo grandes organizaciones en sectores como farmacia, banca, energía o tecnología, entre otras, son capaces de tratar y aprovechar los *big data*"

Más allá de los tecnicismos, en la actualidad sólo las grandes organizaciones en sectores como las farmacéuticas, la banca, grandes empresas de energía o tecnológicas, entre otras, son capaces de permitirse una tecnología tan compleja. Se requiere personal informático muy especializado y con capacidades no sólo informáticas, sino con una alta especialización matemática o estadística, y con competencias analíticas, económicas y de comunicación para no sólo interpretar y dar valor a los datos, sino ser capaz de sintetizarlos y comunicarlos de forma clara, fiable y eficiente.

¿Será necesario para bibliotecarios y profesionales de la información o la comunicación conocer las técnicas para trabajar con datos masivos? No parece que tal circunstancia pueda darse en un horizonte a medio plazo. Los grandes centros de datos en donde se puede trabajar con tecnologías de datos masivos requieren personal con alta cualificación informática y con un perfil muy multidisciplinar –como hemos dicho- y, sobre todo, una infraestructura tecnológica alejada de lo que conocemos. Ni los profesionales de la información están preparados ni las escuelas o facultades de información o comunicación han sabido prever la demanda de formación en campos tan especializados y, a la vez, tan multidisciplinares.

Pero el mundo no son sólo datos masivos o *big data*. Existen muchos conjuntos de datos a los que se puede sacar valor y en donde los profesionales de la información y de la comunicación quizá tengan un gran papel que jugar: los datos abiertos.

3. Datos abiertos

Por datos abiertos se entiende

> "datos que pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona, y que se encuentran sujetos, cuando más, al requerimiento de atribución y de compartirse de la misma manera en que aparecen" (**Dietrich**; **Gray**; **McNamara**; **Poikola**; **Pollock**; **Tait**; **Zijlstra**, n.d.).

Abierto no significa gratis, sino a un coste razonable o proporcional a su valor. Reutilizables significa que deben estar disponibles en una forma conveniente para poder agregarlos a otros conjuntos de datos. Y redistribuibles significa que dichos datos deben ser provistos de licencias o términos de acuerdo que permitan usarlos, sin otras restricciones comerciales o de ningún otro tipo.

Desde el punto de vista técnico, Tim Berners-Lee añadió a sus notas técnicas sobre *linked data* un sistema de marcas basado en estrellas (**Berners-Lee**, 2009) para concienciar sobre la importancia de que los datos sean realmente abiertos y, a ser posible, vinculados:

- 1 estrella: si los datos están disponibles en la Web, en cualquier formato, y con una licencia abierta;
- 2 estrellas: cuando además los datos se ofrecen en un formato estructurado, por ejemplo, en *Excel* en vez de imágenes o pdfs;
- 3 estrellas: como dos estrellas, pero además los datos se ofrecen en un formato no propietario, por ejemplo, ficheros de datos separados por comas (CSV) en vez de *Excel*;
- 4 estrellas: cuando además de todo lo anterior se usan estándares (RDF, y Sparql) para poder identificar cosas y para que otros puedan apuntar hacia ellas; y
- 5 estrellas: los datos vinculados, todo lo anterior y además cuando los datos se enlazan con otros para proporcionar contexto.

Los datos abiertos actuales generalmente son públicos, generados por las administraciones públicas. Por eso muchas veces se les confunde con gubernamentales o como parte del llamado e-gobierno o gobierno electrónico. Que sean públicos no significa que sean muy abiertos, de hecho, muchos documentos y datos gubernamentales en nuestro país son abiertos, pero sólo tienen 1 estrella (mucho pdf, poco reutilizables). Pueden existir también datos abiertos generados por empresas, pero las entidades privadas temen perder cierto control sobre su negocio liberando datos que les afectan, por lo que el grueso

SCIPEDIA
Register for free at https://www.scipedia.com to download the version without the watermark

de los datos abiertos se puede decir que son datos generados por organismos públicos.

Open Knowledge International, antes *Open Knowledge Foundation* (*OKFN*, n.d) distingue 7 tipos de datos abiertos:

1) Culturales: datos sobre trabajos culturales y objetos, por ejemplo, títulos y autores- generalmente recopilados por GLAM (galerías, bibliotecas, archivos y museos);

2) Científicos: se producen como parte de la investigación científica, desde la astronomía a la zoología.

3) Finanzas: sobre cuentas públicas (gastos e ingresos) e información sobre mercados financieros.

4) Estadísticos: producidos por las oficinas estadísticas, como datos del censo o indicadores socioeconómicos clave.

5) Meteorológicos: los muchos tipos de información utilizada para comprender y predecir el tiempo y el clima.

6) Medioambientales: sobre el medio ambiente natural, como presencia y niveles de polución, calidad de ríos y mares...

7) Transporte: horarios, rutas, estadísticas de puntualidad, etc.

Cuando se habla de la ventaja de los datos abiertos a menudo se mencionan tres tipos de beneficios (**Hernández-Pérez**; **García-Moreno**, 2013):

- Transparencia para el buen funcionamiento de las sociedades democráticas, puesto que permite conocer qué hacen los gobiernos. Como ejemplo de transparencia se puede mencionar el proyecto de la *Fundación Civio* "¿Dónde van mis impuestos?" (*Civio*, n.d.) en donde, a partir de los datos de los *Presupuestos generales del Estado* (por cierto, extraídos a partir de ficheros de datos abiertos no muy reutilizables) el ciudadano puede visualizar información sobre gastos e ingresos del Estado.
- Aporte de valor comercial y social, puesto que permite la creación de negocios y servicios innovadores basados en esos datos, como las aplicaciones de muchas ciudades españolas para ofrecernos información sobre si se debe esperar un autobús o la demora hace que convenga más tomar otro medio de transporte. O las aplicaciones con información sobre el tiempo, la calidad del aire, o de las aguas en las playas.
- Participación y compromiso de los ciudadanos puesto que al estar más informados pueden implicarse más en los procesos de toma de decisiones, hacer sugerencias sobre presupuestos participativos o dar ideas sobre el uso que se puede asignar a terrenos sin usar de una comunidad autónoma.

No se trata, por tanto, de trabajar con datos masivos sino de aprender a trabajar con datos abiertos y, preferiblemente, vinculados.

4. Las bibliotecas universitarias y la gestión de datos científicos

El mundo de los datos ha irrumpido con fuerza en las bibliotecas académicas desde que la *OCDE* (2007) publicara un documento en el que instaba a los líderes políticos y científicos a impulsar políticas para facilitar el acceso a los datos de investigaciones financiadas con fondos públicos. Las razones esgrimidas en el informe aludían a que los retos de la humanidad son cada vez más globales (salud, cambio climático, etc.) y que había que aprovechar las nuevas tecnologías para lograr mejorar el retorno de la inversión que hacen las instituciones públicas cuando financian la investigación.

Desde entonces las agencias de financiación de la investigación más importantes del mundo han ido tomando iniciativas para incrementar la puesta a disposición del público de los datos de investigación. Existe un amplio consenso (**Hossain**; **Dwivedi**; **Rana**, 2016) en que la transparencia y la reproducibilidad de las investigaciones son clave para una ciencia abierta y de excelencia. Transparencia en este contexto significa que la metodología publicada en los resultados de investigación ha sido la más adecuada y se aplicó de forma correcta. La reproducibilidad se refiere a la capacidad de los investigadores de aprovechar los datos para confirmar o refutar los resultados, con el mismo conjunto de datos u otros diferentes, agregados o no, que permitan generar nuevos resultados.

La liberación de estos datos abiertos de investigación beneficia a la institución y a la sociedad en general porque:

- acelera la investigación, puesto que se puede ahorrar tiempo en comparar y contrastar datos y resultados, y ayuda a reducir duplicidades o investigaciones redundantes;
- ahorra costes, puesto que grupos pequeños pueden aprovechar los datos de grupos más grandes o viceversa, se pueden agregar datos pequeños; y
- porque comparar y contrastar los datos también supone un ahorro económico y de tiempo. Permite detectar fraudes y que se inicien o se reconsideren líneas de investigación basadas en datos erróneos.

> "La liberación de los datos abiertos de investigación beneficia a la institución y a la sociedad porque acelera la investigación, ahorra costes y permite detectar fraudes y reconsiderar líneas de investigación"

Cualquiera de los informes de prospectiva que tomemos señala la gestión de los datos científicos como uno de los temas que más interés suscita en el mundo de las bibliotecas académicas: la *Association of College & Research Libraries* (*ACRL*, 2014) o el informe *Horizon* (**Johnson**; **Adams-Becker**; **Estrada**; **Freeman**, 2015) señalan la gestión de datos científicos como una de las tendencias que más van a crecer en los próximos tres años. Probablemente recaiga sobre las bibliotecas responder a las presiones de las agencias de financiación sobre los investigadores para que publiquen sus datos de investigación. La UE, por ejemplo, acaba de aprobar una comunicación (*Comisión Europea*, 2016) en la que declara que, a partir de 2017, todos los datos científicos producidos en el marco del *Programa Horizon 2020* deberán estar abiertos por defecto.

SCIPEDIA
Register for free at https://www.scipedia.com to download the version without the watermark

Tabla 1. Cuestiones pendientes para las bibliotecas universitarias

- Realizar guías y protocolos para implementar planes de gestión de datos, como los ya publicados por el *Consorcio Madroño* (*Madroño*, n.d.)	- Orientar y formar a los investigadores sobre dónde y cómo poder encontrar planes de gestión o datos de investigación aplicables a sus investigaciones.
- Orientar y formar a los investigadores sobre la conveniencia de formatos y estándares a utilizar para poder preservar sus datos a largo plazo.	- Orientar y formar a los investigadores sobre dónde depositar (qué repositorios) y cómo (esquemas de metadatos a utilizar en función del área de conocimiento) para que sus datos sean descubribles.
- Revisar sus infraestructuras y sus procesos y decidir si conviene almacenar en repositorios institucionales o en repositorios de cada disciplina, los datos generados en el seno de su institución.	- Establecer políticas que aseguren que los investigadores cumplen con las obligaciones y requisitos (anonimización de los datos, fechas de liberación de datos, etc.) que establecen las agencias de financiación y las autoridades académicas.
- Ayudar a los investigadores a que sus datos sean visibles y puedan ser citados.	- Enlazar los datos de investigación con otros registros académicos (tesis, artículos u otros objetos digitales):

A las bibliotecas universitarias les tocará, sin duda, resolver cuestiones como las indicadas en la tabla 1.

Pasquetto *et al.* (2015) resumen muy bien los ocho ejes principales alrededor de los cuales gira el interés respecto a los datos de investigación (figura 2), que pasamos a comentar:

Las definiciones sobre *open data* (OD) se refieren al concepto de "abierto" en el contexto del mundo de la investigación y de las políticas que deben implementarse para considerar que los datos sean técnica y legalmente abiertos.

Las fuentes de los datos abiertos de investigación (*open research data*) son los estudios que distinguen este tipo de datos de otros como los gubernamentales, y describen las iniciativas (repositorios, revistas, etc.) que tienen lugar en los distintos campos.

Al hacer accesibles los datos de investigación se obtienen beneficios económicos, educativos (para formar a nuevos científicos), sociales –como el apoyo que puede suponer para países menos desarrollados–, y se impulsa la ciencia hecha por los ciudadanos, además de la mencionada, transparencia, reproducibilidad y otros principios relacionados con la ciencia abierta.

Escala de compartición de datos se refiere a: quiénes, cuánto y cómo se comparten, puesto que desde el punto de vista internacional se trata de una práctica que aún está en su infancia. No hay incentivos para publicar datos, aunque ahora se trabaja en el reconocimiento de sus citas (data citation).

Los aspectos sobre la propiedad y las licencias se refieren a la ética y a la privacidad. Los datos y los metadatos pueden revelar información personal muy sensible. Y la propiedad de los datos puede ser un foco de problemas cuando existen conflictos entre la industria privada y la administración pública por la propiedad de los mismos.

Ventajas e inconvenientes de los distintos medios que se utilizan para publicar los datos. Algunos repositorios comerciales garantizan el acceso, pero no la preservación a largo plazo. Los repositorios institucionales son más difíciles de gestionar debido a la diversidad de las áreas de conocimiento, a la falta de cultura entre investigadores. Algunas revistas obligan a depositar en repositorios concretos; otras permiten almacenar los datos en los propios servidores de los grupos de investigación.

El acceso técnico a los datos aborda el tema de las relaciones entre los conjuntos de datos (*datasets*) y otros productos de investigación, desde los resultados a los programas que se han utilizado; desde las bases de datos generadas a partir de la recolección de datos a la metodología que se ha empleado para el análisis o la visualización de dichos datos.

SCIPEDIA
Register for free at https://www.scipedia.com to download the version without the watermark

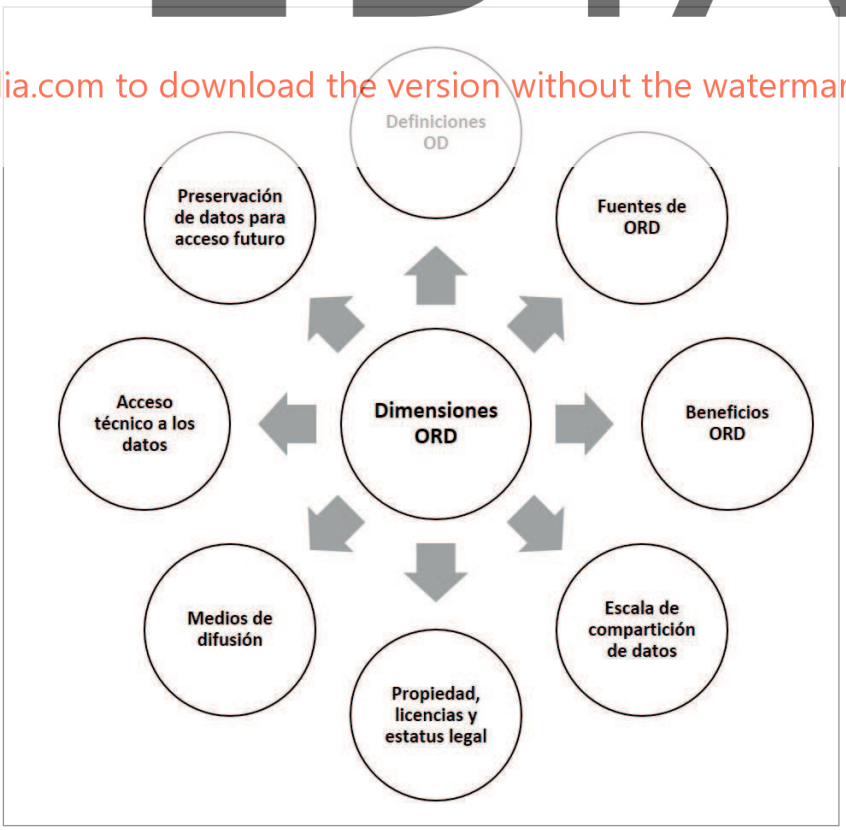

Figura 2. Ejes principales de investigación sobre datos científicos. Basado en **Pasquetto** *et al.* (2015). OD = *Open data*, ORD = *Open research data*

Y sobre la reutilización de los mismos y su capacidad de estar vinculados.

Por último, la preservación de los datos es algo que preocupa especialmente a archiveros y bibliotecarios, quién tendrá la responsabilidad de la gestión de los datos, de su almacenamiento -si se almacenan en repositorios externos-, de su curación, acceso o mantenimiento sostenible. Además del problema de la propia tarea de describir con metadatos estos *datasets*.

5. Bibliotecas públicas y datos abiertos

También las bibliotecas públicas tienen un papel que jugar con los datos abiertos, aunque no sea con el tipo específico de datos de investigación. Las bibliotecas no dejan de innovar para intentar seguir atrayendo a su comunidad de usuarios a su espacio. Un reto difícil considerando que la mayor parte de la población satisface muchas de las necesidades que antes suplía la biblioteca desde prácticamente cualquier sitio con acceso a internet.

Estamos viviendo cómo las bibliotecas se transforman arquitectónicamente, quitando espacio a las estanterías de libros y liberándolos para ordenadores, impresoras 3D o cómodos espacios en donde realizar otro tipo de actividades: desde la alfabetización digital a los cuentacuentos o a talleres de escritura. Se trata, como decía Barbara Striping, presidenta de la *American Library Association* entre 2013 y 2014, de lograr que las bibliotecas

> "transformen su relación con las comunidades y darles el poder para que éstas sean creadoras de información y no solo consumidoras" (**De-Boer**, 2015).

¿En tiempos de sobreabundancia de información, cómo puede la biblioteca pública contribuir a aportar valor social e implicar a los ciudadanos en los procesos de toma de decisiones de la comunidad de usuarios a los que atiende mediante datos abiertos? Un ejemplo, el *Ayuntamiento de Madrid* publica datos diarios de contaminación acústica desde su web de medio ambiente:
http://goo.gl/ieOX2N

Los datos se publican por barrios y por tramo horario (mañana, tarde y noche). En zonas con problemas de contaminación acústica, ¿resulta imaginable que la biblioteca proporcione formación, ayude a descargar los datos y permita hacer estudios más específicos sobre los focos de contaminación para demostrar la elevada contaminación acústica en su área de trabajo y proponer al ayuntamiento algunas iniciativas para reducirla?

Quizá hoy no resulte muy imaginable. No se trata de un reto fácil, las bibliotecas han dispuesto durante mucho tiempo de millones de datos que no se han sabido explotar. Con honrosas excepciones, como el proyecto *Library analytics toolkit* (**Dulin**; **Spina**, 2013), la mayor parte de las bibliotecas no han sabido explotar el sistema de recomendación que tan de moda ha puesto *Amazon* ("si usted leyó esto quizá le interese esto otro") u ofrecer a un estudiante de primero de derecho un listado con los títulos de los libros de primer curso de carrera más prestados el año anterior. O mostrar si en los últimos cinco años se compran muchos libros de matemáticas, pero no se presta ninguno. Los datos no han sido, al menos hasta ahora, una gran prioridad.

Existen muchos más ejemplos, datos sobre accidentes con lesiones, con información sobre el lugar, tipo de accidente, horas, días de la semana…, o sugerencias y reclamaciones presentadas por los ciudadanos para proponer mejoras en los servicios o para reclamar o protestar sobre el normal o anormal funcionamiento de los servicios municipales que permitirían demostrar con datos fehacientes que no se hace caso de una sugerencia o reclamación recurrente de los ciudadanos, etc.

El rol de las bibliotecas públicas respecto a los datos abiertos se podría resumir en los aspectos que se indican en la tabla 2.

Ahora sí que resulta inimaginable, existen muchas barreras: falta formación en tratamiento de datos, por eso muchas instituciones académicas se lanzan a reformar o crear nuevos planes de estudio, introduciendo cursos de "ciencia de datos", visualización, estadística o programación básica. Pero si no somos conscientes de que todos, teóricos y profesionales, tenemos que seguir reciclándonos cada día, la ola que trae *Google* y los nuevos hábitos de búsqueda de información acabará con muchas bibliotecas y centros de documentación.

Si en esencia, el papel de las bibliotecas públicas consiste en proporcionar acceso a información y conocimiento a la comunidad a la que sirve. Si su misión es lograr que los ciudadanos a los que sirven tengan un espacio que contribuya a lograr una sociedad más democrática, con menos desigual-

SCIPEDIA
Register for free at https://www.scipedia.com to download the version without the watermark

Tabla 2. Rol de las bibliotecas públicas respecto a los datos abiertos

- Identificar y recopilar los conjuntos de datos (*datasets*) abiertos publicados y que afecten a su comunidad.	- Analizar, visualizar y crear contenidos que pongan en valor los *datasets* y promuevan la concienciación y la participación ciudadana.
- Planificar e identificar los conjuntos de datos necesarios para comprender y reflejar la realidad y los problemas que afecten a sus comunidades.	- Abogar ante las administraciones pertinentes la recolección y publicación de datos que puedan afectar y ser de interés para su comunidad de usuarios.
- Identificar iniciativas de otras bibliotecas y comunidades sobre uso de datos abiertos que podrían ser de aplicación en la comunidad de usuarios a los que atiende.	- Formar a usuarios en la identificación y tratamiento de datos cuantitativos y en la visualización o narrativa de dichos datos.
- Organizar y estimular eventos (*hackathon*, etc.) que ayuden a crear aplicaciones que permitan utilizar estos conjuntos de datos.	- Crear conciencia sobre el valor de los datos y metadatos en el mundo digital y su incidencia en la privacidad y los derechos del individuo.

dades económicas (las de quienes no pueden adquirir libros y otros materiales) y menos desigualdades socio-culturales (las de quienes tienen las habilidades y el conocimiento para tomar decisiones bien fundamentadas), entonces las bibliotecas, además de ofrecer infraestructura y recursos físicos y tecnológicos, deberían intensificar su rol como formadoras en programas de alfabetización: básica, digital, informacional y de datos.

Habrá que esperar aún algunos años para ver cómo van evolucionando las bibliotecas, como espacios físicos y como espacios culturales. De lo que no cabe duda es que más que estanterías con los libros, la biblioteca si quiere sobrevivir tendrá que convertirse en una plataforma (*library as platform*) (**Weinberger**, 2012) en la que los bibliotecarios no sólo deberán salir más allá de las estanterías. Quizá haya llegado el momento de, además de ir a buscar a los usuarios en tantas redes sociales, salir más allá de los muros de la biblioteca para encontrar a los usuarios en sus comunidades.

Si reconocemos que entre las misiones principales de las bibliotecas están: promover la igualdad, el acceso y la participación abierta entre la ciudadanía, entonces tendremos que abordar las estrategias que sean necesarias para que las bibliotecas puedan ser centros de participación ciudadana. Centros en los que las bibliotecas y los bibliotecarios juegan el papel de conectores, de activistas sociales fortaleciendo las relaciones sociales de su comunidad.

> "Tendremos que abordar las estrategias que sean necesarias para que las bibliotecas puedan ser centros de participación ciudadana en los que los bibliotecarios juegan el papel de conectores, de activistas sociales."

La necesidad de redefinir los espacios en las bibliotecas debe servir como estrategia para además de seguir siendo un lugar perfecto para la lectura y el estudio, un lugar de juegos, un lugar en el que construir cosas (*makerspaces*) y un lugar en el que se generan contenidos de interés para la comunidad, desde apps a visualizaciones o narrativas sobre su entorno. Y los profesionales de la información deben convencerse de que los usuarios no son sólo consumidores, que además se han ido a otras tiendas, sino creadores y ciudadanos que pueden ver en las bibliotecas un espacio comunitario de innovación y desarrollo social.

The Aspen Institute (2014), en un interesante informe sobre una nueva visión de las bibliotecas públicas lo resume muy bien: es necesario redefinir el papel de las bibliotecas como instituciones que inspiren el aprendizaje, hagan crecer el capital social de sus poblaciones y sean capaces de crear oportunidades entre la comunidad de usuarios a las que atiende. Entre la lista de estrategias que cita para lograr estos objetivos menciona:

- conectar los recursos de otras agencias o bibliotecas a la biblioteca como plataforma, más que reinventar la rueda o ir siempre sola;
- apoyar fuertemente el rol del aprendizaje para que tenga lugar en los espacios físicos o a través de la biblioteca como plataforma; o
- hacer del acceso a la información gubernamental un modelo para curar datos abiertos.

Y es que: las bibliotecas públicas o cambian o sucumben a que sus espacios sean ocupados.

6. *Linked open data* (LOD): datos abiertos vinculados

Si decíamos al principio que cada día la web evoluciona desde una web de documentos hacia una web de datos no podemos dejar de resaltar la importancia que tiene en el mundo de la web semántica, en el que las bibliotecas han jugado un papel crucial, el concepto de datos abiertos vinculados. Como decían **Zeinstra** y **Keller** (2011), "los datos pueden estar abiertos, pero no vinculados. Los datos pueden estar vinculados, pero no abiertos".

Aclaremos antes que a veces se utilizan datos enlazados y datos vinculados como sinónimos. De hecho, cuando se habla de uno de los proyectos más importantes de datos vinculados en España, el de la *Biblioteca Nacional* (*datos.bne.es*) se habla de "Datos enlazados en la *BNE*". Para Xavier Agenjo, la diferencia entre el término enlazado y vinculado está en relación con el tipo de enlace entre dos objetos digitales. En su opinión, que compartimos, si se conectan mediante un URL (*uniform resource locator*, web sintáctica), debería hablarse de datos enlazados; y debería utilizarse datos vinculados

> "para aquellos recursos que tengan una relación mediante una URI (*uniform resource identifier*), que es lo que ocurre en la web semántica" (**Agenjo**, 2012).

El concepto de datos vinculados hace referencia a que los datos, por ejemplo, las diferentes palabras de un documento web que nombra a personas, están vinculados a través de técnicas de web semántica, especialmente RDF (*resource description framework*). RDF es un conjunto de especificaciones del *World Wide Web Consortium* (*W3C*) cuyo fin es describir recursos mediante triples o tripletas en forma de expresiones sujeto-predicado-objeto. Ejemplo: La afirmación 'El cielo tiene color azul' se puede representar en RDF con una tripleta compuesta por un sujeto (cielo), un predicado (un atributo, color) y un valor (azul).

Así es como se está pasando de la web de documentos, con simples textos y palabras, a la web de datos en la que se describe la función de cada palabra. La tecnología LOD permite identificar dentro de un documento el dato que se refiere a una persona, a un lugar, a un acontecimiento, etc., y además describirlo, representarlo y vincularlo con la información que se tiene de esa persona en otros sitios web, sin importar el idioma o el sitio web o URL físico en el que se encuentre.

Existe mucho publicado sobre datos vinculados a partir de las ideas iniciales de **Berners-Lee** (2009). Las reglas para trabajar con datos vinculados son básicamente cuatro:

SCIPEDIA
Register for free at https://www.scipedia.com to download the version without the watermark

> "1) Utilizar URIs como nombres para las cosas (las URIs buenas para la web semántica son las que no cambian);
>
> 2) Utilizar URIs http (URIs desreferenciables) para que la gente pueda ver esos nombres;
>
> 3) Cuando alguien busca una URI, proporcionar información útil usando los estándares apropiados (RDF, Sparql); y
>
> 4) Incluir enlaces a otras URIs, de tal forma que se pueda recuperar más información" (**Méndez**; **Greenberg**, 2012).

Las aplicaciones en el mundo de las bibliotecas son múltiples. **Agenjo** y **Hernández** (2016) acaban de publicar un resumen de un informe del año 2016 de *Library technology reports* sobre el estado actual de los datos vinculados en bibliotecas, archivos y museos en el mundo. Estos autores, muy activos en *Europeana* y otros proyectos españoles (**Hernández**, 2015) aprovechan la ocasión para, con buen criterio, mencionar algunos de los proyectos españoles realmente *linked open data*, como el ya mencionado de la *BNE* y algunos otros.

Antes ya mencionamos el informe *NMC Horizon, Library edition 2015* (**Johnson**; **Adams-Becker**; **Estrada**; **Freeman**, 2015) para señalar que una de las tendencias que destacaba en el mundo de las bibliotecas académicas era la gestión de datos científicos. Pues bien, otra de estas tendencias es "la web semántica y los datos vinculados" y augura que *Bibframe*, un modelo de datos basado en los principios de datos vinculados, será el próximo estándar para asignar y gestionar metadatos bibliográficos sustituyendo a MARC. Permitirá visibilizar mucho más la información contenida en los catálogos de las bibliotecas. En el mismo informe se afirma que:

> "los motores de búsqueda más populares apenas tocan el 10% de internet puesto que el 90% restante son sitios web que no se indexan porque la mayor parte de estos datos residen en bibliotecas, archivos o museos que no pueden ser buscados o detrás de áreas seguras que no pueden ser accedidas por los robots" (p. 42).

Los datos demuestran que las bibliotecas, nacionales, académicas y públicas, están aumentando la exposición de datos vinculados (**Smith-Yoshimura**, 2016). Se publican fundamentalmente datos bibliográficos, pero también ficheros de autoridades, colecciones digitales, datos geográficos, datos sobre objetos de museos y algunos más. Y aunque existen muchas barreras: alta curva de aprendizaje que supone trabajar con datos vinculados; inconsistencia de los datos con los que a veces se trabaja, o, simplemente, falta de recursos, la tendencia a trabajar con datos vinculados no es una moda pasajera y será del todo necesaria si queremos "competir" para que nuestros recursos sean visibles en la Web.

6. Conclusiones

Los ciudadanos han encontrado en *Google* y las redes sociales una alternativa válida para satisfacer sus necesidades de información. Ante esta situación las bibliotecas se están viendo forzadas a redefinir y reorientar sus espacios, sus procesos y sus servicios para adaptarse y poder ofrecer sus recursos de forma virtual, además de seguir prestando sus infraestructuras físicas y tecnológicas para no perder la esencia de servicio cultural que siempre han tenido.

Las bibliotecas necesitan salir de las estanterías y lograr un mayor compromiso social con los ciudadanos a los que atienden. Con cada vez menos recursos, tienen que demostrar el valor que aportan a la sociedad. En la carrera por aumentar y ofrecer nuevos servicios es necesario una mayor implicación con los investigadores y con los ciudadanos. Los datos abiertos pueden ser una oportunidad para ello y eso requiere seleccionar sobre qué servicios se debe poner el acento. Los *big data* están fuera del alcance de la mayor parte de las bibliotecas, públicas y académicas, pero los datos abiertos pueden ser una gran oportunidad para impulsar el papel de formadoras que tradicionalmente han jugado las bibliotecas.

En las bibliotecas académicas, concienciando, formando y gestionando, desde planes de gestión de datos de investigación hasta el descubrimiento y preservación de los mismos, para ayudar a los investigadores a satisfacer la presión de las agencias de financiación para liberar los datos de investigación y contribuir así a mejorar y acelerar una verdadera ciencia abierta. En las bibliotecas públicas, aprovechando las posibilidades que ofrecen los datos abiertos para generar contenidos que atañen muy directamente a sus comunidades y propiciando espacios de innovación y de participación ciudadana. Para todas ellas, exponer sus datos en forma de datos abiertos vinculados será una de las mejores formas de poner en valor sus fondos y hacerse más visible ante la sociedad. Los datos abiertos y especialmente los datos abiertos vinculados entran dentro de la categoría de estos "nuevos" servicios, sin excluir a los demás. Renovarse o morir. No hacer nada no puede ser una opción.

Nota

1. *Apache Hadoop* es una plataforma de software de código abierto para el almacenamiento y procesamiento distribuidos de grandes conjuntos de datos en clusters de ordenadores. Todos los módulos de *Hadoop* están diseñados en el supuesto de que los fallos de hardware son comunes y deben gestionarse automáticamente dentro de *Hadoop*.

2. *NoSQL* (no SQL o no relacional) en un programa para elaborar bases de datos con un sistema de almacenamiento y recuperación que no se basa en relaciones tabulares como las utilizadas en las bases de datos relacionales. Han existido desde la década de 1960, pero no se las llamó *NoSQL* hasta que a principios del siglo XXI aumentó su popularidad por necesidades de las empresas web 2.0 como *Facebook*, *Google* y *Amazon*. Se utilizan cada vez más en *big data* y aplicaciones web en tiempo real.

7. Bibliografía

Agenjo, Xavier (2012). "Tómatelo con filología: no es lo mismo enlazado que vinculado". *Archivos de la lista de distribución de IweTel*.
https://listserv.rediris.es/cgi-bin/wa?A2=IWETEL;a9862cc8.1209D

Agenjo, Xavier; **Hernández, Francisca** (2016). "El estado de los datos vinculados en bibliotecas en 2015". *Blok de bid*.
http://www.ub.edu/blokdebid/es/content/el-estado-de-los-datos-vinculados-en-bibliotecas-en-2015

The Aspen Institute (2014). "Rising to the challenge: Re-envisioning public libraries", Dialogue on public libraries, 80 pp.

SCIPEDIA
Register for free at https://www.scipedia.com to download the version without the watermark

http://csreports.aspeninstitute.org/documents/AspenLibrariesReport.pdf

Berners-Lee, Tim (2009). "Linked data - Design issues". *https://www.w3.org/DesignIssues/LinkedData.html*

Civio, Fundación Ciudadana (n.d). "¿Dónde van mis impuestos?". *http://www.dondevanmisimpuestos.es*

Comisión Europea (2014). "Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las regiones. *Hacia una economía de los datos próspera. COM (2014). 442 final*". *http://goo.gl/OVHTC0*

Comisión Europea (2016). "Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones. European cloud initiative - *Building a competitive data and knowledge economy in Europe. COM (2016). 178 final*". *http://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=COM:2016:178:FIN&rid=2*

ACRL (2014). "Top trends in academic libraries: A review of the trends and issues affecting academic libraries in higher education". *College & research libraries news*, v. 75, n. 6, pp. 294-302. *http://crln.acrl.org/content/75/6/294*

De-Boer, Jeroen (2015). "The business case of FryskLab, Europe's first mobile library FabLab". *Library Hi Tech*, v. 33, n. 4, pp. 505-518. *http://dx.doi.org/10.1108/LHT-06-2015-0059*

Dietrich, Daniel; **Gray, Jonathan**; **McNamara, Tim**; **Poikola, Antti**; **Pollock, Rufus**; **Tait, Julian**; **Zijlstra, Ton** (n.d). "¿Qué son los datos abiertos?". *Open data handbook*. *http://opendatahandbook.org/guide/es/what-is-open-data*

Furlo, Kirk; **Kozak, Caleb** (2014). "Using analytics to improve Harvard Library Lab". *https://osc.hul.harvard.edu/liblab/projects/library-analytics-toolkit*

Hernández, Francisca (2015). "Los cambios en *Europeana data model* 5.2.5 y 5.2.6". *Blok de bid*. *http://www.ub.edu/blokdebid/es/content/los-cambios-en-europeana-data-model-525-y-526*

Hernández-Pérez, Tony; **García-Moreno, María-Antonia** (2013). "Datos abiertos y repositorios de datos: nuevo reto para los bibliotecarios". *El profesional de la información*, v. 22, n. 3, pp. 259-263. *http://dx.doi.org/10.3145/epi.2013.may.10*

Hossain, Mohammad A.; **Dwivedi, Yogesh K.**; **Rana, Nripendra P.** (2016). "State-of-the-art in open data research: Insights from existing literature and a research agenda". *Journal of organizational computing and electronic commerce*, v. 26, n. 1-2, pp. 14-40. *http://dx.doi.org/10.1080/10919392.2015.1124007*

IBM (2016). "What is big data?". https://www-*01.ibm.com/software/data/bigdata/what-is-big-data.html*

Johnson, Larry; **Adams-Becker, Samantha**; **Estrada, Victoria**; **Freeman, Alex** (2015). *NMC Horizon report: 2015 Library edition*, 60 pp. *http://cdn.nmc.org/media/2015-nmc-horizon-report-library-EN.pdf*

Madroño, Consorcio (n.d). *Pagoda: PlAn de GestiÓn de DAtos del Consorcio Madroño*. *http://www.consorciomadrono.es/pagoda*

Méndez, Eva; **Greenberg, Jane** (2012). "Datos enlazados para vocabularios abiertos: marco global de HIVE". *El profesional de la información*, v. 21, n. 3, pp. 236-244. *http://dx.doi.org/10.3145/epi.2012.may.03*

Naciones Unidas (2014). *A world that counts: Mobilising the data revolution for sustainable development*, 32 pp.. *http://goo.gl/94ehcc*

Negredo, Samuel; **Vara-Miguel, Alfonso**; **Amoedo, Avelino** (2016). *Digital news report .es 2016: Cambios decisivos en el consumo de noticias digitales*", 84 pp. *http://www.digitalnewsreport.es*

NIST (2015). *NIST big data interoperability framework: Vol. 1, Definitions*. Big Data Public Working Group, Definitions and Taxonomies Subgroup. *http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1.pdf*

Obama, Barack (2009). "Transparency and open government". *The White House*. *https://www.whitehouse.gov/node/698*

OECD (2007). *OECD Principles and guidelines for access to research data from public funding*. OECD. *http://goo.gl/Rh2Bhb*

OECD (2015). *Data-driven innovation*, OECD Publishing, ISBN: 978 92 64 22934 1. *http://goo.gl/pt5cWe*

OKFN (n.d). *Open knowledge: What is open? https://okfn.org/opendata*

Pasquetto, Irene V.; **Sands, Ashley E.**; **Borgman, Christine L.** (2015). "Exploring openness in data and science: What is "open," to whom, when, and why?". *Procs of the Association for Information Science and Technology*, v. 52, n. 1, pp. 1-2. *http://dx.doi.org/10.1002/pra2.2015.1450520100141*

Rowley, Jennifer (2007). "The wisdom hierarchy: representations of the DIKW hierarchy". *Journal of information science*, v. 33, n. 2, pp. 163-180. *http://dx.doi.org/10.1177/0165551506070706*

Smith-Yoshimura, Karen (2016). "Analysis of international linked data survey for implementers". *D-Lib magazine*, v. 22, n. 7/8. *http://dx.doi.org/10.1045/july2016-smith-yoshimura*

Weinberger, David (2012). "Library as platform". *Library journal*. Sept. 4th *http://lj.libraryjournal.com/2012/09/future-of-libraries/by-david-weinberger*

Zeinstra, Maarten; **Keller, Paul** (2011). *Open linked data and Europeana*. *http://goo.gl/aKYRCc*

SCIPEDIA
Register for free at https://www.scipedia.com to download the version without the watermark